PATENT
Attorney Docket 4232.1US

CERTIFICATE OF MAILING

Express Mail Label Number: EL 740547394 US

Date of Deposit: September 26, 2001

Person making Deposit: Daniel Thatcher

APPLICATION FOR LETTERS PATENT

for

SMAD-INTERACTING POLYPEPTIDES AND THEIR USE

Inventors:
Kristin Verschueren
Jacques Remacle
Danny Huylebroeck

Attorney:
Allen C. Turner
Registration No. 33,041
TRASKBRITT, PC
P.O. Box 2550
Salt Lake City, Utah 84110
(801) 532-1922

TITLE OF THE INVENTION

SMAD-INTERACTING POLYPEPTIDES AND THEIR USE

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is a divisional application of co-pending U. S. patent application no. 09/449,285, filed on 24 November 1999, now U. S. Patent __________, which itself claims priority from pending application PCT/EP98/03193 filed on 28 May 1998 designating the United States of America, which itself claims priority from European Patent Application EP 97201645.5 filed on 2 June 1997.

REFERENCE TO A "SEQUENCE LISTING"

The computer readable form of the sequence listing in this application is identical with that filed in U. S. patent application no. 09/449,285, filed 24 November 1999. In accordance with 37 CFR § 1.821(e), please use the last-filed computer readable form filed in that application as the computer readable form for the instant application. The paper copy of the instant application is identical with the computer readable copy filed for application no. 09/449,285.

TECHNICAL FIELD

[0002] The present invention relates to SMAD - interacting polypeptides ("SIP's") such as cofactors for SMAD proteins and the use thereof.

BACKGROUND

[0003] The development from a single cell to a fully organized organism is a complex process wherein cell division and differentiation are involved. Certain proteins play a central role in this process. These proteins are divided into different families of which the transforming growth factor β ("TGF-β") family of ligands, their serine/threonine kinase ("STK") receptors and their signalling components are undoubtedly key regulatory polypeptides. Members of the TGF-b superfamily have been documented to play crucial roles in early developmental events such as mesoderm formation and gastrulation, but also at later stages in processes such as neurogenesis, organogenesis, apoptosis and establishment of left-right asymmetry. In addition, TGF-b ligands and components of their signal transduction pathway have been identified as putative tumor suppressors in the adult organism..

[0004] Recently, "SMAD proteins" have been identified as downstream targets of the STK receptors (Massagué,1996, *Cell*, 85, p. 947-950). These SMAD proteins are signal transducers which become phosphorylated by activated type I receptors and thereupon accumulate in the nucleus where they may be involved in transcriptional activation. SMAD proteins comprise a family of at least 5 subgroups which show high cross-species homology. They are generally proteins of about 450 amino acids (50-60kDa) with highly conserved N-terminal and C-terminal domains, linked by a variable, proline-rich, middle region. On the basis of experiments carried out in cell lines or in *Xenopus* embryos, it has been suggested that the subgroups define distinct signalling pathways: SMAD1 mediates BMP2/4 pathways, while SMAD2 and SMAD3 act in TGF-b / activin signal transduction cascades. It has been demonstrated that these SMADs act in a complex with SMAD4 (dpc-4) to elicit certain activin, bone morphogenetic protein (BMP) or TGF-b responses (Lagna et al., 1996, *Nature*, 383, p.832-836 and Zhang et al., 1996, *Nature*, 383, p.168-172).

[0005] SMAD proteins have a three-domain structure and their highly conserved carboxyl domain (C-domain) is necessary and sufficient for SMAD function in the nucleus. The concept that this domain of SMAD proteins might interact with transcription factors in order to regulate transcription of target genes has previously been put forth (Meersseman et al, 1997, *Mech.Dev.*, 61, p.127-140). This hypothesis has been supported by the recent identification of a new winged-helix transcription factor ("FAST1") which forms an activin-dependent complex with SMAD2 and binds to an activin responsive element in the Mix-2 promoter (Chen et al., *Nature* 383, p. 691-696, 1996). However, cofactors for SMAD proteins other than FAST 1 have not yet been identified.

[0006] Beyond the determination of the mechanism of activation of STK receptors and SMAD, and the heteromerization of the latter, little is known about other downstream components in the signal transduction machinery. Thus, understanding how cells respond to TGF-b related ligands remains a crucial central question in this field.

[0007] In order to clearly demonstrate that SMAD proteins might have a function in transcriptional regulation -either directly or indirectly- it is necessary to identify putative co-factors of SMAD proteins, response elements in target genes for these SMAD proteins and/or co-factors, and to investigate the ligand-dependency of these activities.

[0008] To understand those interactions molecular and developmental biology research on (i) functional aspects of the ligands, receptors and signaling components (in particular members of the SMAD family), in embryogenesis and disease, (ii) structure-function analysis of the ligands and the receptors, (iii) the elucidation of signal transduction, (iv) the identification of cofactors for SMAD (related) proteins, and (v) ligand-responsive genes in cultured cell and the *Drosophila*, amphibian, fish and murine embryo are all of utmost importance.

DISCLOSURE OF THE INVENTION

[0009] We have found that by carrying out a two hybrid screening assay, SMAD interacting protein(s) are obtainable where SMAD C-domain fused to a DNA-binding domain as "bait" and a vertebrate cDNA library as "prey" respectively are used. It is evident to those of skill in the art that other appropriate cDNA libraries can be used as well. By using, for example, SMAD1 C-domain fused to GAL4 DNA-binding domain and a mouse embryo cDNA as bait and prey respectively, a partial SMAD4 and other SMAD-interacting protein (SIP) cDNAs, including SIP1, were obtained.

[0010] Surprisingly, it has been found that at least four SMAD interacting proteins thus obtained contain a DNA binding zinc finger domain. One of these proteins, SIP1, is a novel member of the family of zinc finger/homeodomain proteins containing d-crystallin enhancer binding protein and certain *Drosophila* zfh-1, the former of which has been identified as a DNA-binding repressor. It has been shown that one DNA binding domain of SIP1 (the C-terminal zinc finger cluster or SIP1_{czf}) binds to E2 box regulatory sequences and to the *Brachyury* protein binding site. It has been demonstrated in cells that SIP1 interferes with E2 box and *Brachyury*-mediated transcription activation. SIP1 fails to interact with full-size SMAD in yeast. We have shown for the first time that SMAD proteins can interact with a DNA-binding repressor and, as such, appear to be directly involved in TGF-ß ligand-controlled repression of target genes which are involved in the strict regulation of normal early development.

[0011] In summary, characteristics of SIP 1 include the following:

a) it fails to interact with full size XSMAD1 in yeast,

b) it is a new member of the family of zinc finger/homeodomain proteins including δ-crystallin enhancer binding protein and/or *Drosophila* zfh-1,

c) SIP1_{czf} binds to E2 box sites,

d) SIP1_{czf} binds to the Brachyury protein binding site,

e) it interferes with Brachyury-mediated transcription activation in cells, and

f) it interacts with C-domain of SMAD 1, 2 and/or 5.

[0012] As used herein, "E2 box sites" means a -CACCTG- regulatory conserved nucleotide sequence which contains the binding site CACCT for δ-crystallin enhancer binding proteins as described in Sekido et al, 1996, *Gene*, 173, p.227-232. These E2 box sites are known targets for important basic helix-loop-helix (bHLH) factors such as MyoD , a transcription factor in embryogenesis and myogenesis.

[0013] So, the SIP1 according to the invention (a zinc finger/homeodomain protein) binds to specific sites in the promoter region of a number of genes which are relevant for the immune response and early embryogenesis and as such may be involved in transcriptional regulation of important differentiation genes in significant biological processes such as cell growth and differentiation, embryogenesis, and abnormal cell growth including cancer.

[0014] The invention also includes an isolated nucleic acid sequence including the nucleotide sequence as provided in SEQ ID NO: 1 coding for a SMAD interacting protein or a functional fragment thereof.

[0015] Furthermore, a recombinant expression vector including the isolated nucleic acid sequence (in sense or anti-sense orientation) operably linked to a suitable control sequence belongs to the present invention and cells transfected or transduced with a recombinant expression vector as well.

[0016] Another aspect of the invention is a polypeptide including the amino acid sequence according to SEQ ID NO: 2 or a functional fragment thereof. The present invention also includes variants or homologues of amino acids enclosed in the disclosed polypeptides wherein the amino acids are modified and/or substituted by other amino acids obvious for a person skilled in the art. For example, post-expression modifications of the polypeptide such as phosphorylations are not excluded from the scope of the current invention.

[0017] A pharmaceutical composition including the previously identified nucleic acid(s) or a pharmaceutical composition including the polypeptide(s) are another aspect of the invention. The nucleic acid and/or polypeptide according to the invention can be optionally used for appropriate gene therapy purposes.

[0018] In addition, a method for diagnosing, prognosis and/or follow-up of a disease or disorder by using the nucleic acid(s) according to the invention or by using the polypeptide(s) also form an important aspect of the current invention. Furthermore, in the method for diagnosing, prognosis and/or follow-up of a disease or disorder an antibody, directed against a polypeptide or fragment thereof according to the current invention, can also be conveniently used. As used herein, the term "antibody" refers, without limitation, to preferably purified polyclonal antibodies or monoclonal antibodies, altered antibodies, univalent antibodies, Fab proteins, single domain antibodies or chimeric antibodies. In many cases, the binding phenomena of antibodies to antigens is equivalent to other ligand/anti-ligand binding.

[0019] A diagnostic kit including a nucleic acid(s) sequence and/or a polypeptide(s) or antibodies directed against the polypeptide or fragment thereof according to the invention for performing previously identified method for diagnosing a disease or disorder clearly belong to the invention as well.

[0020] Diseases or disorders in this respect are, for instance, related to cancer, malformation, immune or neural diseases, or bone metabolism related diseases or disorders. In addition a disease affecting organs like skin, lung, kidney, pancreas, stomach, gonad, muscle or intestine can be diagnosed as well using the diagnostic kit according to the invention.

[0021] Using the nucleic acid sequences of the invention as a basis, oligomers of approximately 8 nucleotides or more can be prepared, either by excision or synthetically, which hybridize for instance with a sequence coding for SIP or a functional part thereof and are thus useful in identification of SIP in diseased individuals. These so-called "probes" are of a length which allows the detection of unique sequences of the compound to detect or determine by hybridization as defined above. While 6-8 nucleotides may be a workable length, sequences of about 10 -12 nucleotides are preferred, and about 20 nucleotides appears optimal. The nucleotide sequence may be labelled for example with a radioactive compound, biotin, enzyme, dye stuff or metal sol, fluorescent or chemiluminescent compound. The probes can be packaged into diagnostic kits. Diagnostic kits include the probe nucleotide sequence, which may be labeled; alternatively, the probe may be unlabeled and the ingredients for labelling may be included in the kit in separate containers so that the probe can optionally be labeled. The kit may also contain other suitably packaged reagents and materials needed for the particular hybridization protocol, for example, standards, wash buffers, as well as instructions for conducting the test.

[0022] The diagnostic kit may include an antibody directed to a polypeptide or fragment thereof according to the invention in order to set up an immunoassay. Design of the immunoassay is subject to a great deal of variation, and the variety of these are known in the art. Immunoassays may be based, for example, upon competition, or direct reaction, or sandwich type assays.

[0023] An important aspect of the present invention is the development of a method of screening for compounds (chemically synthesized or available from natural sources) which affect the interaction between SMAD and SIP's having the current knowledge of the SMAD interacting polypeptides (so called SIP's such as SIP1 or SIP2 as specifically disclosed herein).

[0024] A transgenic animal harbouring the nucleic acid(s) according to the invention in its genome also belong to the scope of this invention. The transgenic animal can be used for testing medicaments and therapy models as well. As used herein, a transgenic animal means a non-human animal which has incorporated a foreign gene (called transgene) into its genome. Because this gene is present in germ line tissues, it is passed from parent to offspring establishing lines of transgenic animals from a first founder animal. As such, transgenic animals are recognized as specific species variants or strains, following the introduction and integration of new gene(s) into their genome. The term "transgenic" has been extended to chimeric or "knockout" animals in which gene(s), or part of genes, have been selectively disrupted or removed from the host genome.

[0025] It will be appreciated that when a nucleic acid construct is introduced into an animal to make it transgenic, the nucleic acid may not necessarily remain in the form as introduced.

[0026] As used herein, "offspring" means any product of the mating of the transgenic animal whether or not with another transgenic animal, provided that the offspring carries the transgene.

[0027] Depending on the purpose of the gene transfer study, transgenes can be grouped into three main functional types: *gain-of-function*, *reporter function* and *loss-of-function*.

[0028] The *gain-of-function* transgenes are designed to add new functions to the transgenic individuals or to facilitate the identification of the transgenic individuals if the genes are expressed properly (including in some cell types only) in the transgenic individuals.

[0029] The *reporter gene function* is commonly used to identify the success of a gene transfer effort. Bacterial chloramphenicol acetyltransferase ("CAT"), b-galactosidase or luciferase genes fused to functional promoters represent one type of *reporter function* transgene.

[0030] The *loss-of-function* transgenes are constructed for interfering with the expression of host genes. These genes might encode an antisense RNA to interfere with the post-transcriptional process or translation of endogenous mRNAs. Alternatively, these genes might encode a catalytic RNA (a ribozyme) that can cleave specific mRNAs and thereby cancel the production of the normal gene product.

[0031] Optionally, loss of function transgenes can also be obtained by over-expression of dominant-negative variants that interfere with activity of the endogenous protein or by targeted inactivation of a gene , or parts of a gene, in which usually (at least a part of) the DNA is deleted and replaced with foreign DNA by homologous recombination. This foreign DNA usually contains an expression cassette for a selectable marker and/or reporter.

[0032] The invention also includes a SMAD interacting protein characterized in that:

a) it interacts with full size XSMAD1 in yeast,

b) it is a member of a family of proteins which contain a cluster of 5 CCCH-type zinc fingers including *Drosophila* "Clipper" and Zebrafish "No arches",

c) it binds single or double stranded DNA,

d) it has an RNase activity, and

e) it interacts with C-domain of SMAD1, 2 and/or 5.

[0033] The invention also includes a method for post-transcriptional regulation of gene expression by members of the TGF-b superfamily by manipulation or modulation of the interaction between SMAD function and/or activity and mRNA stability.

BRIEF DESCRIPTION OF THE FIGURE

[0034] FIG. 1 shows that the XSMAD1 C-domain interacts with SIP1 in mammalian cells and deletion of the 51 amino acid ("aa") long SBD (SMAD binding domain) in SIP1 abolishes the interaction. COS-1 cells were transiently transfected with expression constructs encoding N-terminally myc-tagged SIP1 and a GST-XSMAD1 C-domain fusion protein. The latter was purified from cell extracts using gluthatione-sepharose beads. Purified proteins were visualized after SDS-PAGE and Western blotting using anti-GST antibody (Pharmacia), (Panel A, slim arrow). Myc-tagged SIP1 protein was co-purified with GST-XSMAD1 C-domain fusion protein, as was shown by Western blotting of the same material using anti-myc monoclonal antibody (Santa Cruz)(Panel C, lane one, thick arrow). Deletion of the 51 aa long SBD in SIP1 abolished this interaction (panel C, lane 2). Note that the amounts of purified GST-XSMAD1 C-domain protein and levels of expression of both SIP1 (wild type and SIP1del SBD) proteins in total cell extracts were comparable (compare lanes 1 and 2 in panel A and B).

DETAILED DESCRIPTION OF THE INVENTION

[0035] A two hybrid screening assay for use with the invention may be performed as generally described by Chien et al., *PNAS*, 88, p.9578-9582. (1991).

[0036] The polypeptide or fragments thereof included within the invention are not necessarily translated from the nucleic acid sequence according to the invention but may be generated in any manner, including, for example, chemical synthesis or expression in a recombinant expression system. Generally, "polypeptide" refers to a polymer of amino acids, and does not refer to a specific length of the molecule. Thus, linear peptides, cyclic or branched peptides, peptides with non-natural or non-standard amino acids such as D-amino acids, ornithine and the like, oligopeptides and proteins are all included within the definition of polypeptide. The terms "protein" and "polypeptide", as used herein, are generally interchangeable. "Polypeptide" as previously mentioned refers to a polymer of amino acids (amino acid sequence) and does not refer to a specific length of the molecule. Thus, peptides and oligopeptides are included within the definition of polypeptide. This term also includes post-translational modifications of the polypeptide, for example, glycosylations, acetylations, phosphorylations and the like. Included within the definition are, for example, polypeptides containing one or more analogs of an amino acid (including, for example, unnatural amino acids, etc.), polypeptides with substituted linkages, as well as other modifications known in the art, both naturally occurring and non-naturally occurring.

[0037] "Control sequence", as used herein, refers to regulatory DNA sequences which are necessary to affect the expression of coding sequences to which they are ligated. The nature of such control sequences differs depending upon the host organism. In prokaryotes, control sequences generally include promoter, ribosomal binding site, and terminators. In eukaryotes, generally control sequences include promoters, terminators and, in some instances, enhancers, transactivators, transcription factors or 5' and 3' untranslated cDNA sequences. The term "control sequence" is intended to include, at a minimum, all components the presence of which are necessary for expression, and may also include additional advantageous components.

[0038] "Operably linked", as used herein, refers to a juxtaposition wherein the components so described are in a relationship permitting them to function in their intended manner. A control sequence "operably linked" to a coding sequence is ligated in such a way that expression of the coding sequence is achieved under conditions compatible with the control sequences. In case the control sequence is a promoter, it would be obvious to a skilled person to use double-stranded nucleic acid.

[0039] As used herein, "fragment of a sequence" or "part of a sequence" means a truncated sequence of the original sequence referred to. The truncated sequence (nucleic acid or protein sequence) can vary widely in length; the minimum size being a sequence of sufficient size to provide a sequence with at least a comparable function and/or activity of the original sequence referred to, while the maximum size is not critical. In some applications, the maximum size usually is not substantially greater than that required to provide the desired activity and/or function(s) of the original sequence. Typically, the truncated amino acid sequence will range from about 5 to about 60 amino acids in length. More typically, however, the sequence will be a maximum of about 50 amino acids in length, preferably a maximum of about 30 amino acids. It is usually desirable to select sequences of at least about 10, 12 or 15 amino acids, up to a maximum of about 20 or 25 amino acids.

[0040] Furthermore, the current invention is not limited to the exact isolated nucleic acid sequences specifically identified herein, including the nucleotide sequence as mentioned in SEQ ID NO: 1, but also a nucleic acid sequence hybridizing to the nucleotide sequence as provided in SEQ ID NO: 1 or a functional part thereof and encoding a SMAD interacting protein or a functional fragment thereof belongs to the present invention.

[0041] To clarify, as used herein, "hybridization" means conventional hybridization conditions known to the skilled person, preferably appropriate stringent hybridization conditions. Hybridization techniques for determining the complementarity of nucleic acid sequences are known in the art. The stringency of hybridization is determined by a number of factors during hybridization including temperature, ionic strength, length of time and composition of the hybridization buffer. These factors are outlined in, for example, Maniatis et al. (1982) *Molecular Cloning; A laboratory manual* (Cold Spring Harbor Press, Cold Spring Harbor, N.Y.).

[0042] The term "antigen" refers to a polypeptide or group of peptides which include at least one epitope. "Epitope" refers to an antibody binding site usually defined by a polypeptide including 3 amino acids in a spatial conformation which is unique to the epitope, generally an epitope consists of at least 5 such amino acids and more usually of at least 8-10 such amino acids.

[0043] The invention is further explained by the following illustrative EXAMPLES:

EXAMPLES

Example I

Yeast, two-hybrid cloning of SMAD-interacting proteins

[0044] In order to identify cofactors for SMAD1, a two-hybrid screening in yeast was carried out using the *X*SMAD1 C-domain fused to GAL4 DNA-binding domain (GAL4$_{DBD}$) as bait, and a cDNA library from mouse embryo (12.5 dpc) as a source of candidate preys. The GAL4$_{DBD}$-SMAD1 bait protein failed to induce in the reporter yeast strain GAL4-dependent *HIS3* and *LacZ* transcription on its own or in conjunction with an empty prey plasmid. Screening of 4 million yeast transformants identified about 500 colonies expressing *HIS3* and *LacZ*. The colonies displaying a phenotype which was dependent on expression of both the prey and the bait cDNAs, were then characterized. Plasmids were rescued and the prey cDNAs sequenced (SEQ ID NOS: 1-20 of the

Sequence Listing enclosed; for each nucleic acid sequence only one strand is depicted in the Listing). Four of these (th1, th12, th76 and th74 respectively also denominated in this application as SIP1, SIP2, SIP5 and SIP7 respectively) are disclosed in detail (embedded in SEQ ID NOS: 1, 2, 3, 4, 10, and 8 respectively). One (th72= combined SEQ ID NO: 6 and 7) encodes a protein in which the GAL4 transactivation domain (GAL4$_{\text{TAD}}$) is fused in-frame to a partial SMAD4 cDNA, which starts at amino acid (aa) 252 in the proline-rich domain. SMAD4 has been shown to interact with other SMAD proteins, but no SMAD has been picked-up thus far in a two-hybrid screen in yeast, using the C-domain of another SMAD as bait. These data suggest that the N-domain of both interacting SMAD proteins, as well as part of (SMAD4) or the entire (SMAD1) proline-rich domain, is dispensable for heterodimeric interaction between SMAD proteins, at least when using a two-hybrid assay in yeast.

[0045] The cDNA insert of the second positive prey plasmid, th1 (embedded in SEQ ID NO: 1), encodes a protein in which the GAL4$_{\text{TAD}}$-coding sequence is fused in-frame to about a 1.9 kb-long th1 cDNA, which encodes a polypeptide SIP1 (Th1) of 626 aa. Data base searches revealed that SIP1 (Th1) contained a homeodomain-like segment, and represents a novel member of a family of DNA-binding proteins including vertebrate d-crystallin enhancer binding proteins (d-EF1) and *Drosophila* zfh-1. These zinc finger/ homeodomain-containing transcription factors are involved in organogenesis in mesodermal tissues and/or development of the nervous system. The protein encoded by th1 cDNA is a SMAD interacting protein (SIP) and was named SIP1 (TH1).

Example II

SIP1

Characterization of SIP1-SMAD interaction in yeast and *in vitro*

[0046] The binding of SIP1 (TH1) to full-size XSMAD1 and modified C-domains was tested. The latter have either an amino acid substitution (G418S) or a deletion of the last 43 aa (D424-466). The first renders the SMAD homolog in *Drosophila* Mad inactive and abolishes BMP-dependent phosphorylation of SMAD1 in mammalian cells. A truncated Mad, similar to mutant D424-466, causes mutant phenotypes in *Drosophila*, while a similar truncation in SMAD4 (dpc-4) in a loss-of-heterozygosity background is associated with pancreatic carcinomas. SIP1 (TH1) does neither interact with full-size *X*SMAD1, nor with mutant D424-466. The absence of any detectable association of full-size *X*SMAD1 was not due to inefficient expression of the latter in yeast, since one other SMAD-interacting prey (th12) efficiently interacted with the full-length SMAD bait. Lack of association of SIP1 (TH1) with full-size *X*SMAD1 in yeast follows previous suggestions that the activity of the SMAD C-domain is repressed by the N-domain, and that this repression is eliminated in mammalian cells by incoming BMP signals. The G418S mutation in the C-domain of SMAD 1 does not abolish interaction with SIP1, suggesting that this mutation affects another aspect of SMAD1 function. The ability of the full-size G418S SMAD protein to become functional by activated receptor STK activity may thus be affected, but not the ability of the G418S C-domain to interact with downstream targets. This indicates that activation of SMAD is a prerequisite for and precedes interaction with targets such as SIP1. The deletion in mutant D424-466 includes three conserved and functionally important serines at the C-terminus of SMAD which are direct targets for phosphorylation by the activated type I STK receptor.

[0047] The C-domains of SMAD1 and SMAD2 induce ventral or dorsal mesoderm, respectively, when over-expressed individually in *Xenopus* embryos, despite their very high degree of sequence conservation. Very recently, SMAD5 has been shown to induce ventral fates in the *Xenopus* embryo. To investigate whether the striking differences in biological activity of SMAD1, -5 and SMAD2 could be due to distinct interactions with cofactors, the ability of SIP1 (TH1) protein to interact with the C-domains of SMAD1, -5 and SMAD2 in a yeast two-hybrid assay was tested. SIP1 (TH1) was found to interact in yeast with the C-domain of all three SMAD members. Then the interaction of SIP1 with different SMAD C-domains *in vitro* was investigated, using glutathione-S-transferase ("GST") pull-down assays. GST-SMAD fusion proteins were produced in *E. Coli* and coupled to glutathione-Sepharose beads. An unrelated GST fusion protein and unfused GST were used as negative controls. Radio-labeled, epitope-tagged SIP1 protein was successfully produced in mammalian cells using a vaccinia virus (T7VV)-based system. Using GST-SMAD beads, this SIP1 protein was pulled down from cell lysates, and its identity was confirmed by Western blotting. Again, as in yeast, it was found that SIP1 is a common binding protein for different SMAD C-domains, suggesting that SIP1 might mediate common responses of cells to different members of the TGF-ß superfamily. Alternatively, SMAD proteins may have different affinities for SIP1 *in vivo*, or other mechanisms might determine the specificity, if any, of SMAD-SIP1 interaction.

Example III

SIP1 is a new member of zinc finger/homeodomain proteins of the dEF-1 family

[0048] Additional SIP1 open reading frame sequences were obtained by a combination of cDNA library screening with 5'RACE-PCR. The screening yielded a 3.2 kb-long SIP1 cDNA (tw6), which overlaps partially with th1 cDNA. The open reading frame of SIP1 protein encodes 944 aa (SEQ ID NO: 2), and showed homology to certain regions in d-EF1, ZEB, AREB6, BZP and zfh-1 proteins, and strikingly similar organisation of putative functional domains. Like these proteins, SIP1 contains two zinc finger clusters separated by a homeodomain and a glutamic acid-rich domain. Detailed comparisons reveal that SIP1 is a novel and divergent member of the two-handed zinc finger/homeodomain proteins. As in d-EF1, three of the five residues that are conserved in helix 3 and 4 of all canonical homeodomains are not present in SIP1. SIP1 (Th1) which contains the homeodomain but lacks the C-terminal zinc finger cluster and glutamic acid-rich sequence, interacts with SMAD. This interaction is maintained upon removal of the homeodomain-like domain, indicating that a segment encoding aa 44-236 of SIP1 (numbering according to SEQ ID NO: 2) is sufficient for interaction with SMAD. To narrow this domain further down, progressive deletion mutants, starting from the N-terminus, as well as the C-terminus of this 193 aa region were made. Progressive 20 aa deletion constructs were generated by PCR. Two restriction sites (5' end SmaI site, 3' end XhoI site) were built in to allow cloning of amplified sequences in the yeast two hybrid bait vector pACT2 (Clontech). An extensive two hybrid experiment was performed with these so-called SBD mutant constructs as a prey and the XSMAD1 C-domain as bait. The mutant SBD constructs that encoded aa 166-236 (of SEQ ID NO: 2) or aa 44-216 were still able to interact with the bait plasmid, whereas mutant constructs encoding aa 186-236 or aa 44-196 could not interact with the bait. In this way, the smallest domain that still interacts with the XSMAD1 C-domain was defined as a 51 aa domain encompassing aa 166-216 of SEQ ID NO: 2.

[0049] The amino acid sequence of the SBD, necessary for the interaction with SMAD, thus is (depicted in one-letter code):

QHLGVGMEAPLLGFPTMNSNLSEVQKVLQIVDNTVSRQKMDCKTEDISKLK (SEQ ID NO: 21)

[0050] Deletion of an additional 20 aa at the N-or C-terminal end of this region disrupted the SMAD binding activity. Subsequently, this 51aa region was deleted in the context of SIP1 protein, again using a PCR based approach, generating an NcoI restriction site at the position of the deletion. This SIP1DSBD51 was not able to interact with the SMAD C-domain any longer, as assayed by a "mammalian pull down assay". In these experiments, SIP1, myc-tagged at its N-terminal end was expressed in COS-1 cells together with a GST-*X*SMAD1 C-domain fusion protein. Myc-SIP1 protein was co-purified from cell extracts with the GST-*X*SMAD1 C-domain fusion protein using gluthatione-sepharose beads, as was demonstrated by Western blotting using anti-myc antibody. Deletion of the 51 aa in SIP1 abolished the interaction, as detected in this assay, with the *X*SMAD1 C-domain. (*See*, FIG. 1).

Example IV

Analysis of the DNA-binding activity of the C-terminal zinc finger cluster of SIP1.

[0051] d-EF1 is a repressor that regulate the enhancer activity of certain genes. This repressor binds to the E2 box sequence (5'-CACCTG) which is also a binding site for a subgroup of basic helix-loop-helix (bHLH) activators (Sekido et al., 1994, *Mol.Cell.Biol.*,14, p. 5692-5700). Interestingly, the CACCT sequence which has been shown to bind d-EF1 is also part of the consensus binding site for Bra protein. It has been proposed that cell type-specific gene expression is accomplished by competitive binding to CACCT sequences between repressors and activators. δ-EF1 mediated repression could be the primary mechanism for silencing the IgH enhancer in non-B cells. d-EF1 is also present in B-cells, but is counteracted by E2A, a bHLH factor specific for B-cells. Similarly, d-EF1 represses the Igk enhancer where it competes for binding with bHLH factor E47.

[0052] The C-terminal zinc finger cluster of dEF-1 is responsible for binding to E2 box sequences and for competition with activators. Considering the high sequence similarities in this region between SIP1 and d-EF1, it was decided to test first whether both proteins have similar DNA binding specificities, using gel retardation assays. Therefore, the DNA-binding properties of the C-terminal zinc finger cluster of SIP1 (named SIP1$_{CZF}$) was analyzed. SIP1$_{CZF}$ was efficiently produced in and purified from *E. coli* as a short GST fusion protein. Larger GST-SIP1 fusion proteins were subject to proteolytic degradation in *E. coli* .

[0053] Purified GST-SIP1$_{CZF}$ was shown to bind to the E2 box of the IgH kE2 enhancer. A mutation of this site (Mut1), which was shown previously to affect the binding of the bHLH factor E47 but not d-EF1, did not affect binding of SIP1$_{CZF}$. Two other mutations in this kE2 site (Mut2 and Mut4, respectively) have been shown to abolish binding of d-EF1 (Sekido *et al.*, 1994) and did so in the case of SIP1$_{CZF}$. In addition, also the binding of SIP1$_{CZF}$ to the Nil-2A binding site of the interleukin-2 promoter, the Bra protein binding site and the AREB6 binding site were demonstrated. The specificity of the binding of SIP1$_{CZF}$ to the Bra binding site was further demonstrated in competition experiments. Binding of SIP1$_{CZF}$ to this site was competed by excess unlabeled Bra binding site probe, while kE2 wild type probe competes, albeit less efficiently than its variant Mut1, which is a very strong competitor. kE2-Mut2 and kE2-Mut4 failed to compete, as did the GATA-2 probe, while the AREB6 site competed very efficiently. From these experiments, it can be concluded that GST-SIP1$_{CZF}$ fusion protein displays the same DNA binding specificity as other GST fusion proteins made with the CZF region of d-EF1 and related proteins (Sekido *et al.*, 1994). In addition, it was demonstrated for the first time that SIP1 binds specifically to regulatory sequences that are also target sites for Bra. This may be the case for the other d-EF1-related proteins as well and these may interfere with Bra-dependent gene activation *in vivo*.

[0054] Analyses were done to sites recognized by the bHLH factor MyoD. MyoD has been shown to activate transcription from the muscle creatine kinase ("MCK") promoter by binding to E2 box sequences (Weintraub et al., 1994, *Genes Dev.*,8, p.2203-2211; Katagiri *et al.*, 1997, *Exp.Cell Res.* 230, p. 342-351). Interestingly, d-EF1 has also been demonstrated to repress MyoD-dependent activation of the MCK enhancer, as well as myogenesis in 10T½ cells, and this is thought to involve E2 boxes (Sekido *et al.*, 1994). In addition, TGF-ß and BMP-2 have been reported to down-regulate the activity of muscle-specific promoters, and this inhibitory effect is mediated by E2 boxes (Katagiri *et al.*, 1997). The latter are present in the regulatory regions of many muscle-specific genes, are required for muscle-specific expression, and are optimally recognized by heterodimers between myogenic bHLH proteins (of the MyoD family) and of widely expressed factors like E47. SIP1$_{CZF}$ was able to bind to a probe that encompasses the MCK enhancer E2 box and this complex was competed by the E2 box oligonucleotide and by other SIP1 binding sites. In addition, a point mutation within this E2 box that is similar to the previously used kE2-Mut4 site also abolished binding of SIP1$_{czf}$. These results confirm that SIP1$_{czf}$ binds to the E2 box of the MCK promoter. SIP1, as SMAD-interacting and MCK E2 box binding protein, may therefore represent the factor that mediates the TGF-ß and BMP repression of the MyoD-regulated MCK promoter (Katagiri *et al.*, 1997).

Example V

SIP1 is a BMP-dependent repressor of Bra activator

[0055] The experiments have demonstrated that SIP1$_{CZF}$ binds to the Bra protein binding site, IL-2 promoter, and to E2 boxes, the latter being implicated in BMP or TGF-ß-mediated repression of muscle-specific genes. These observations prompted therefore to test whether SIP1 (as SIP1$_{TW6}$) is a BMP-regulated repressor. A reporter plasmid containing a SIP1 binding site (the Bra protein binding site) fused to the luciferase gene was constructed. COS cells, maintained in low serum (0.2%) medium during the transfection, were used in subsequent transient transfection experiments since they have been documented to express BMP receptors and support signalling (Hoodless *et al.*, 1996,Cell, 85, p.489-500). It was found in the experiment that SIP1$_{TW6}$ is not able to change the transactivation activity of Bra protein via the Bra binding site. In addition, no transactivation of this reporter plasmid by SIP1$_{TW6}$ could be detected in the presence of 10% or 0.2% serum, and in the absence of Bra expression vector.

[0056] Therefore, identical experiments were carried out in which the cells were exposed to BMP-4. SIP1$_{TW6}$ repressed the Bra-mediated activation of the reporter. It does this in a dose-dependent fashion (amount of SIP1$_{TW6}$ plasmid, concentration of BMP-4). Total repression has not been obtained in this type of experiment, because the transfected COS cells were exposed only after 24 hours to BMP-4. Consequently, luciferase mRNA and protein accumulate during the first 24 hours of the experiment as the result of Brachyury activity. The conclusion from these experiments clearly shows that SIP1 is a repressor of Bra activator, and its activity as repressor is detected only in the presence of BMP. It is important that SIP1 has not been found to be an activator of transcription via Bra target sites. This is interesting, since the presence in d-EF1-like proteins of a polyglutamic acid-rich stretch (which is also present in SIP1_{TW6} used here) has led previously to the speculation that these repressors might act as transcriptional activators as well. In particular, AREB6 has been shown to bind to the promoter of the housekeeping gene Na,K- ATPase a-1 and to repress gene expression dependent on cell type and on the context of the binding site (Watanabe *et al.*, 1993, *J. Biochem.*,114, p. 849-855).

Example VI

SIP1 mRNA expression in mice

[0057] Northern analysis demonstrated the presence of a major SIP1 6 kb mRNA in the embryo and several tissues of adult mice, with very weak expression in liver and testis. A minor 9 kb-long transcript is also detected, which is however present in the 7 dpc embryo. *In situ* hybridization documented SIP1 transcription in the 7.5 dpc embryo in the extra-embryonic and embryonic mesoderm. The gene is weakly expressed in embryonic ectoderm. In the 8.5 dpc embryo, very strong expression is seen in extra-embryonic mesoderm (blood islands), neuroepithelium and neural tube, the first and second branchial arches, the optic eminence, and predominantly posterior presomitic mesoderm. Weaker but significant expression is detected in somites and notochord. Between day 8.5 and 9.5, this pattern extends clearly to the trigeminal and facio-acoustic neural crest tissue. Around mid-gestation, the SIP1 gene is expressed in the dorsal root ganglia, spinal cord, trigeminal ganglion, the ventricular zone of the frontal cortex, kidney mesenchyme, non-epithelial cells of duodenum and mid-gut, pancreatic primordium, urogenital ridge and gonads, the lower jaw and the snout region, cartilage primordium in the humerus region, the primordium of the clavicle and the segmental pre-cartilage sclerotome-derived condensations along the vertebral axis. SIP1 mRNA can also be detected in the palatal shelf, lung mesenchyme, stomach and inferior ganglion of vagus nerve. In addition, primer extension analysis has demonstrated the presence of SIP1 mRNA in embryonic stem cells. It is striking that the expression of SIP1 in the 8.5 dpc embryo in the blood islands and presomitic mesoderm coincides with tissues affected in BMP-4 knockout mice, which have been shown to die between 6.5 and 9.5 dpc with a variable phenotype. These surviving till later stages of development showed disorganized posterior structures and a reduction in extra-embryonic mesoderm, including blood islands (Winnier *et al.*, 1995, *Genes Dev.*, 9, 2105-2116).

[0058] The mRNA expression of d-EF1 proteins has been documented as well. In mouse, d-EF1 mRNA has been detected in mesodermal tissues such as notochord, somites and nephrotomes, and in other sites such as the nervous system and the lens in the embryo (Funahashi *et al.*, 1993, Development, 119, p.433-446). In adult hamster, d-EF1 mRNA has been detected in the cells of the endocrine pancreas, anterior pituitary and central nervous system (Franklin *et al.*, 1994, *Mol.Cell.Biol.*,14, p. 6773-6788). The majority of these d-EF1 and SIP1 expression sites overlap with sites where the restricted expression pattern of certain type I STK receptors (such as ALK-4/ActR-IA, and ALK-6/BMPR-IB) has been documented (Verschueren *et al.*, 1995, Mech.Dev.,52, p.109-123).

Example VII

SIP2

Characterization of SIP2

[0059] SIP2 was picked up initially as a two hybrid clone of 1052 base pairs ("bp") (th12) that shows interaction in yeast with SMAD1, 2 and 5 C-terminal domains and full-size SMAD1. Using GST-pull down experiments (as described for SIP1) also an interaction with SMAD1, 2 and 5 C-terminal domains *in vitro* have been demonstrated.

a) SIP2 full length sequence

[0060] Th12 showed high homology to a partial cDNA (KIAA0150) isolated from the human myoloblast cell line KG1. However, this human cDNA is +/- 2 kb longer at the 3' end of th12. Using this human cDNA, an EST library was screened and mouse EST were detected homologous to the 3'end of KIAA0150 cDNA. Primers were designed based on th12 sequence and the mouse EST found to amplify a cDNA that contains the stop codon at the 3'end. 5' sequences encompassing the start codon was obtained using 5'RACE-PCR .

[0061] Gene bank accession numbers for the mentioned EST clones used to complete the SIP2 open reading frame: Human KIAA0150 ; D63484 and Mouse EST sequence; Soares mouse p3NMF19.5; W82188.

[0062] Primers used to reconstitute SIP2 open reading frame:
based on th12 sequence: F3th12F (forward primer) 5'-cggcggcagatacgcctcctgca (SEQ ID NO: 22)
based on EST sequence: th12mouse1 (reverse primer) 5'-caggagcagttgtgggtagagccttcatc (SEQ ID NO: 23)

[0063] Primers used for 5'-race; all are reverse primers derived from th12 sequence
1: 5'-ctggactgagctggacctgtctctccagtac (SEQ ID NO: 24)
2 : 5'-cacaagggagtatttcttgcgccacgaagg (SEQ ID NO: 25)
3: 5'-gccatggtgtgaggagaagc (SEQ ID NO: 26)

[0064] The full size SIP2 deduced from the assembly of these sequences contains 950 amino acids as depicted in SEQ ID NO. 4, while the nucleotide sequence is depicted in SEQ ID NO. 3.

b) SIP2 sequence homologies

[0065] SIP2 contains a domain encompassing 5 CCCH type zinc fingers. This domain was found in other protein such as Clipper in *Drosophila*, No Arches in Zebrafish and CPSF in mammals. No Arches is essential for development of the branchial arches in Zebrafish and CPSF is involved in transcription termination and polyadenylation. The domain containing the 5 CCCH in Clipper was shown to have an EndoRNase activity (see below).

c) SIP2 CCCH domain has an RNAse activity

[0066] The domain containing the 5 CCCH -type zinc fingers of SIP2 was fused to GST and the fusion protein was purified from *E. coli.* This fusion protein displays a RNAse activity when incubated with labelled RNA produced *in vitro*. In addition, it has been shown that this fusion protein was able to bind single stranded DNA.

[0067] In more detail, GST fusion proteins of SIP2 5xCCCH; PLAG1 (an unrelated zinc finger protein), SIP1$_{CZF}$ (C-terminal zinc finger cluster of SIP1) and th1 (SIP1 partial polypeptide isolated in the two-hybrid screening), and cytoplasmatic tail of CD40 were produced in *E. coli* and purified using glutathione sepharose beads. Three ^{35}S labelled substrates, previously used to demonstrate the RNAse activity of Clipper, a related protein from *Drosophila* (Bai, C. and Tolias P.P. 1996, cleavage of RNA Hairpins Mediated by a Developmentally Regulated CCCH Zinc Finger Protein. *Mol Cell. Biol.* 16: 6661-6667) were produced by *in vitro* transcription. The RNA cleavage reactions with purified GST fusion proteins were performed in the presence of RNAsin (blocking RNAseA activity). Equal aliquots of each reaction were taken out at time points 1', 7', 15', 30', 60'. Degradation products were separated on a denaturing polyacrylamide gel and visualized by autoradiography. These experiments demonstrated that GST-SIP2 5XCCCH has an RNAse activity and degrades all tested substrates, while GST-PLAG1, GST-CD40, GST-SIP1$_{CZF}$ and GST-th1 do not have this activity.

[0068] Interaction between th12 (partial SIP2 polypeptide) and SMAD C-domains in GST pull down experiments.

[0069] C-domains of *Xenopus (X)*SMAD1 and mouse SMAD2 and 5 were produced in *E. coli* as fusion proteins with gluthatione S-transferase and coupled to gluthatione beads. An unrelated GST-fusion protein (GST-CD40 cytoplasmatic mail) and GST itself were used as negative controls.

[0070] Th12 protein, provided with an HA-tag at its N-terminal end, was produced in Hela cells using the T7 vaccinia virus expression system and metabolically labelled. Expression of Th12 was confirmed by immune precipitation with HA antibody, followed by SDS-page and autoradiography. Th12 protein is produced as a ± 50 kd protein. Cell extracts prepared from Hela cells expressing this protein were mixed with GST-SMAD C-domain beads in GST pull down buffer and incubated overnight at 4° C. The beads were then washed four times in the same buffer, the bound proteins eluted in Laemmli sample buffer and separated by SDS-PAGE. "Pulled down" th12 protein was visualized by Western blotting , using HA antibody. These experiments demonstrate that th12 is efficiently pulled down by GST-SMAD C-domain beads, and not by GST-CD40 or GST alone.

Conclusion on SIP2

[0071] SIP2 is a SMAD interacting protein that contains a RNAse activity. The finding that SMADs interact with potential RNAses provides an unexpected link between the TGF-b signal transduction and mRNA stabilisation.

Example VIII

SIP5

Characterization of SIP5

[0072] One contiguous open reading frame is fused in frame to the GAL4 transactivating domain in the two hybrid vector pACT-2 (Clontech). This represents a partial cDNA, since no in frame translational stop codon is present. The sequence has no significant homology to anything in the database, but displays a region of high homology with following EST clones:

[0073] Mouse: accession numbers: AA212269 (Stratagene mouse melanoma); AA215020 (Stratagene mouse melanoma), AA794832 (Knowles Solter mouse 2 c) and Human: accession numbers AA830033, AA827054, AA687275, AA505145, AA371063.

[0074] Analysis of interaction of the SIP5 prey protein with different bait proteins (which are described in the data section obtained with SIP1) in a yeast two hybrid assay is as follows:

Empty bait vector pGBT9	-
Full length XSMAD1	+
XSMAD1 C-domain	+
XSMAD1 C-domain with G418S substitution	+
Mouse SMAD2 C-domain	+
Mouse SMAD5 C-domain	+
Lamin (pLAM; Clontech)	-

[0075] SIP5 partial protein encoded by above described cDNA also interacts with XSMAD1, mouse SMAD2 and 5 C-domains in vitro as analysed by the GST pull down assay (previously described for SIP1 and SIP2). Briefly, the partial SIP5 protein was tagged with a myc tag at its C-terminal end and expressed in COS-1 cells. GST-SMAD C-domain fusion proteins, GST-CD40 cytoplasmatic tail and GST alone were expressed in *E. coli* and coupled to glutathione sepharose beads. These beads were subsequently used to pull down partial SIP5 protein from COS cell lysates, as was demonstrated after SDS-PAGE of pulled down proteins followed by Western blotting using anti myc antibody. In this assay, SIP5 was pulled down by GST-XSMAD1, 2 and 5 C-domains, but not by GSTalone or GST-CD40. A partial, but coding, nucleic acid sequence for SIP5 is depicted in SEQ ID NO: 10.

Example IX

SIP7 (Characterization of SIP7)

[0076] One contiguous open reading frame is fused in frame to the GAL4 transactivating domain in the two hybrid vector pACT2. This is a partial clone, since no in frame translational stop codon is present. Part of this clone shows homology to Wnt-7b,accession number: M89802, but the clone seems to be a novel cDNA or a cloning artefact. The homology of the SIP7 cDNA with the known Wnt7-b cDNA starts at nucleotide 390 and extends to nucleotide 846. This corresponds to the nucleotides 74-530 in Wnt7-b coding sequences (with A of the translational start codon considered as nucleotide nr 1). In SIP7 cDNA this region of homology is preceded by a sequence that shows no homology to anything in the database. It is not clear whether the SIP7 cDNA is for example a new Wnt7-b transcript or whether it is a scrambled clone as a result of the fusion of two cDNAs during generation of the cDNA library.

[0077] Analysis of the interaction of the SIP7 prey protein with different bait proteins in a yeast two hybrid assay can be summarized as follows:

PGBT9	-
Full length XSMAD1	-
XSMAD1 C-domain	+
XSMAD1 C-domain, G418S	+
XSMAD1 C-domain del aa 424-466	-
XSMAD1 N-terminal domain	-
Mouse SMAD2 C-domain	+
Mouse SMAD5 C-domain	+
Lamin (pLAM)	-

[0078] SIP7 partial protein encoded by above described cDNA also interacts with XSMAD1, mouse SMAD2 and 5 C-domains in vitro as analysed by the GST pull down assay, as described above for SIP5. In this assay, N-terminally myc-tagged SIP7 protein was specifically pulled down by GST-XSMAD1, 2 and 5 C-domains, but not by GSTalone or GST-CD40. A partial, but coding, nucleic acid sequence for SIP7 is depicted in SEQ ID NO: 8.

General description of the methods used

Plasmids and DNA manipulations

[0079] Mouse SMAD1 and SMAD2 cDNAs used in this study were identified by low stringency screening of oligo-dT primed lEx*lox* cDNA library made from 12 dpc mouse embryos (Novagen), using SMAD5 (MLP1.2 clone as described in Meersseman et al., 1997, *Mech.Dev.*,61, p.127-140) as a probe. The same library was used to screen for full-size SIP1, and yielded lExTW6. The tw6 cDNA was 3.6 kb long, and overlapped with th1 cDNA, but contained additional 3'-coding sequences including an in-frame stop codon. Additional 5' sequences were obtained by 5' RACE using the Gibco-BRL 5' RACE kit.

[0080] *X*SMAD1 full-size and C-domain bait plasmids were constructed using previously described *Eco*RI-*Xho*I inserts(Meersseman et al.,1997, Mech.Dev.,61, p.127-140), and cloned between the *Eco*RI and *Sal*I sites of the bait vector pGBT-9 (Clontech), such that in-frame fusions with GAL4$_{DBD}$ were obtained. Similar bait plasmids with mouse SMAD1, SMAD2 and SMAD5 were generated by amplifying the respective cDNA fragments encoding the C-domain using Pfu polymerase (Stratagene) and primers with *Eco*RI and *Xho*I sites. The G418S *X*SMAD1 C-domain was generated by oligonucleotide-directed mutagenesis (Biorad).

[0081] To generate in-frame fusions of SMAD C-domains with GST, the same SMAD fragments were cloned in pGEX-5X-1 (Pharmacia). The phage T7 promoter-based SIP1 (TH1) construct for use in the T7VV system was generated by partial restriction of the th1 prey cDNA with *Bgl*II, followed by restriction with *Sal*I, such that SIP1 (TH1) was lifted out of the prey vector along with an in-frame translational start codon, an HA-epitope tag of the flu virus, and a stop codon. This fragment was cloned into pGEM-3Z (Promega) for use in the T7VV system. A similar strategy was used to clone SIP2 (th12) into pGEM-3Z.

[0082] PolyA$^+$ RNA from 12.5 dpc mouse embryos was obtained with OLIGOTEX-dT (Qiagen). Randomly primed cDNA was prepared using the SUPERSCRIPT CHOICE SYSTEM (Gibco-BRL). cDNA was ligated to an excess of Sfi double-stranded adaptors containing *Stu*I and *Bam*HI sites. To facilitate cloning of the cDNAs, the prey plasmid pAct (Clontech) was modified to generate pAct/Sfi-Sfi. Restriction of this plasmid with *Sfi* generates sticky ends which are not complementary, such that self-ligation of the vector is prevented upon cDNA cloning. A library containing 3.6 X 10^6 independent recombinant clones with an average insert size of 1,100 bp was obtained.

Synthesis of SIP1 and GST pull-down experiments

[0083] Expression of SIP1 (TH1) and SIP2 (TH12) in mammalian cells with the T7VV system and the preparation of the cell lysates were as described previously (Verschueren, K et al.,1995, *Mech.Dev.*,52, p.109-123).

[0084] GST fusion proteins were expressed in *E. coli* (strain BL21) and purified on gluthathione-Sepharose beads (Pharmacia). The beads were washed first four times with PBS supplemented with protease inhibitors, and then mixed with 50 μl of lysate (prepared from T7VV-infected SIP1-expressing mammalian cells) in 1 ml of GST buffer (50 mM Tris-HCl pH 7.5, 120 mM NaCl, 2 mM EDTA, 0.1% (v/v) NP-40, and protease inhibitors). They were mixed at 4°C for 16 hours. Unbound proteins were removed by washing the beads four times with GST buffer. Bound proteins were harvested by boiling in sample buffer, and resolved by SDS-PAGE. Separated proteins were visualized using autoradiography or immunodetection after Western blotting; using anti-HA monoclonal antibody (12CA5) and alkaline phosphatase-conjugated anti-mouse 2ary antibody (Amersham).

EMSA (electrophoretic mobility shift assay)

[0085] The sequence of the kE2 WT and mutated kE2 oligonucleotides are identical as disclosed in Sekido et al; (1994, *Mol.Cell.Biol.*,14, p. 5692-5700). The sequence of the AREB6 oligonucleotide was obtained from Ikeda et al;(1995, *Eur.J.Biochem,* 233, p. 73-82). IL2 oligonucleotide is depicted in Williams et al;(1991, *Science*, 254, p.1791-1794).

[0086] The sequence of Brachyury binding site is 5'-TGACACCTAGGTGTGAATT-3' (SEQ ID NO: 27). The negative control GATA2 oligonucleotide sequences originated from the endothelin promoter (Dorfman et al; 1992, J.Biol.Chem., 267, p. 1279-1285). Double stranded oligonucleotides were labelled with polynucleotide kinase and ^{32}P g-ATP and purified from a 15% polyacrylamide gel. Gel retardation assays were performed according to Sekido et al; (1994, Mol.Cell.Biol.,14, p. 5692-5700).

[0087] RESULTS OF TWO HYBRID SCREENING (XSMAD1 C-domain bait versus 12.5 dpc mouse embryo library; 600.000 recombinant clones screened in 4x 10^6 yeasts).

[0088] SIP 1 - Three independent clones isolated (th1, th88 and th94)

- Zinc-finger-homeodomain protein
- Homology to dEF-1 (see above)
- Interactions in yeast:

XSMAD1 C-domain bait	+
Empty bait	-
Lamin	-
XSMAD1 full length	-
XSMAD1 N-domain	-
mSMAD1 C-domain	+
mSMAD2 C-domain	+
mSMAD5 C-domain	+
XSMAD1 C-domain del 424-466	-
XSMAD1 C-domain G418S	+

* Interaction with C-domain of XSMAD1 and mSMADs confirmed in vitro using GST-pull downs and co-immunoprecipitations

* Extended clone (TW6) isolated through library screening using th1 sequences as a probe

* C-terminal TW6 zinc-finger cluster binds to E2 box sequences (cfr dEF-1), Brachyury T binding site, Brachyury promoter sequences

SIP2 (also called clone TH12)- Three independent clones isolated (th12,th73,th93)

Highly homologous to KIAA0150 gene product, isolated from the myeloblast cell line KG1(Ref: Nagase et al. 1995; *DNA Res* 2 (4) 167-174.

- Interactions in yeast:

XSMAD1 C-domain bait	+
Empty bait	-
Lamin	-
XSMAD1 full length	+
XSMAD1 N-domain	ND
mSMAD1 C-domain	+
mSMAD2 C-domain	+
mSMAD5 C-domain	+
XSMAD1 C-domain del 424-466	-
XSMAD1 C-domain G418S	+

TH60 - Two independent clones isolated (th60 and th77)

- Zinc finger protein

homology to snail (transcriptional repressor) and to ATBF1 (complex homeodomain zinc finger protein)

- Interactions in yeast:

XSMAD1 C-domain bait	+
Empty bait	-
Lamin	-

TH72 - One clone isolated

- Encodes a partial DPC-4 (SMAD4) cDNA (see above)

- Interactions in yeast:

XSMAD1 C-domain bait	+++
Empty bait	-
Lamin	-
XSMAD1 full length	ND
XSMAD1 N-domain	-
mSMAD1 C-domain	+++
mSMAD2 C-domain	ND
mSMAD5 C-domain	+++
XSMAD1 C-domain del 424-466	-
XSMAD1 C-domain G418S	+

SIP5 (also called clone th76).

Analysis of interaction of the SIP5 prey protein with different bait proteins (which are described in the data section obtained with SIP1) in a yeast two hybrid assay can be summarized as follows

Empty bait vector pGBT9	-
Full length XSMAD1	+
XSMAD1 C-domain	+
XSMAD1 C-domain G418S	+
Mouse SMAD2 C-domain	+
Mouse SMAD5 C-domain	+
Lamin (pLAM; Clontech)	-

SIP7 (also called clone th74)

Analysis of the interaction of the SIP7 prey protein with different bait proteins in a yeast two hybrid assay can be summarized as follows:

PGBT9	-
Full length XSMAD1	-
XSMAD1 C-domain	+
XSMAD1 C-domain, G418S	+
XSMAD1 C-domain del aa 424-466	-
XSMAD1 N-terminal domain	-
Mouse SMAD2 C-domain	+
Mouse SMAD5 C-domain	+
Lamin (pLAM)	-

The following clones have been investigated less extensively. They are considered as "true positives" because they interact with the XSMAD1 C-domain bait and not with the empty bait (*i.e.*, GAL-4 DBD alone)

TH75: -Three independent clones isolated (th75, th83, th89)

-Partial aa sequences do not show significant homology to proteins in the public databases

- Interactions in yeast:

XSMAD1 C-domain bait	+++
Empty bait	-

TH92: -Zinc finger protein

-homology to KUP

TH79, TH86, TH90, : Partial sequences do not display significant homology to any protein sequence in the public databases.

Clones available in the sequence listing as conversion table from clone notation to sequence listing notation

SIP 1 nucleotide sequence	= SEQ ID NO: 1
SIP 1 amino acid sequence	= SEQ ID NO: 2
SIP 2 nucleotide sequence	= SEQ ID NO: 3
SIP 2 amino acid sequence	= SEQ ID NO: 4
TH60(TH77)	= SEQ ID NO: 5
TH72 (DPC4 or SMAD4)	= SEQ ID NO: 6
TH72\R	= SEQ ID NO: 7
SIP 7(th74)	= SEQ ID NO: 8
TH75F(TH83F,TH89F)	= SEQ ID NO: 9
SIP 5(th76)	= SEQ ID NO: 10
TH79F	= SEQ ID NO: 11
TH79R	= SEQ ID NO: 12
TH83R	= SEQ ID NO: 13
TH86F	= SEQ ID NO: 14
TH86R	= SEQ ID NO: 15
TH89=TH75R	= SEQ ID NO: 16
TH90F	= SEQ ID NO: 17
TH90R	= SEQ ID NO: 18
TH92F	= SEQ ID NO: 19
TH92R	= SEQ ID NO: 20